Efficient RGB-T Tracking via Cross-Modality Distillation

Tianlu Zhang 1 Hongyuan Guo 1 Qiang Jiao 1 Qiang Zhang 1* Jungong Han 2,3
1 School of Mechano-Electronic Engineering, Xidian University, China
2 Department of Computer Science, the University of Sheffield, UK.
3 Centre for Machine Intelligence, the University of Sheffield, UK.
{tianluzhang, hyg}@stu.xidian.edu.cn, {qzhang, qjiao}@xidian.edu.cn, jungonghan77@gmail.com

Abstract

Most current RGB-T trackers adopt a two-stream structure to extract unimodal RGB and thermal features and complex fusion strategies to achieve multi-modal feature fusion, which require a huge number of parameters, thus hindering their real-life applications. On the other hand, a compact RGB-T tracker may be computationally efficient but encounter non-negligible performance degradation, due to the weakening of feature representation ability. To remedy this situation, a cross-modality distillation framework is presented to bridge the performance gap between a compact tracker and a powerful tracker. Specifically, a specific-common feature distillation module is proposed to transform the modality-common information as well as the modality-specific information from a deeper two-stream network to a shallower single-stream network. In addition, a multi-path selection distillation module is proposed to instruct a simple fusion module to learn more accurate multi-modal information from a well-designed fusion mechanism by using multiple paths. We validate the effectiveness of our method with extensive experiments on three RGB-T benchmarks, which achieves state-of-the-art performance but consumes much less computational resources.

1. Introduction

RGB-T tracking is the task of estimating the state of an arbitrary target in each frame of an RGB-T video sequence [35]. Due to the affordability of thermal infrared (TIR) sensors, RGB-T tracking draws more and more research interest.

As shown in Fig. 1 (a), most existing RGB-T tracking models first adopt a two-stream structure to extract multi-level unimodal RGB and TIR features, respectively, and then employ elaborate-designed multi-modal feature fusion modules to exploit complementary information within the

---

*Corresponding author.
As pointed out by many previous works [9, 17], the shallower layers of unimodal features usually contain abundant low-level spatial details, which are usually modality-dependent. Differently, the deeper layers of unimodal features often contain many high-level semantic cues, which tend to be strongly modality-consistent. The student model uses a compact single-stream network to extract both RGB features and TIR features, which not only lacks the ability to extract modality-specific information in the shallower layers, but also insufficiently explores the modality-common information in the deeper layers. These interesting observations inspire us to design a Specific-common Feature Distillation (SCFD) module, which transforms the modality-specific information as well as the modality-common information from a two-stream deeper network to a single-stream shallower network.

Second, in the stage of multi-modal feature fusion, the complex multi-modal feature fusion modules in the teacher model show great advantages in various scenarios, while the simple fusion strategies in the student model are usually effective in some specific scenarios. It is difficult for a student model with a single simple fusion strategy to learn more effective complementary information mining capabilities from a complex teacher model due to the large feature differences. Therefore, we design a fusion module with multiple simple fusion strategies in the student model, denoted as Multi-path Selection Distillation (MPSD) module. In the process of learning from the teacher model, the student model can adaptively combine different types of fusion features to make up for the lack of complementary information mining capabilities of a single simple fusion strategy.

Finally, in the stage of target state estimation, with the weakening of the feature representation ability of the student model, the discriminative ability of the tracker for distractors is also reduced. For that, we further present a Hard-focused Response Distillation (HFDR) module to improve the student model’s discriminative ability by alleviating the problem of data imbalance between the targets and the backgrounds, which employs the response maps generated by the teacher model to instruct the student to focus on distinguishing targets from hard negative samples.

As shown in Fig. 2, each of our proposed modules continuously reduces the performance gap between the student model and the teacher model without increasing the number of parameters obviously. To sum up, our work improves an RGB-T tracker dramatically because of the following two contributions:

- A Cross-Modality Distillation (CMD) framework is presented to bridge the performance gap between a compact student model and a powerful teacher model through three stages, i.e., unimodal feature extraction, multi-modal feature fusion and target state estimation, as shown in Fig. 1 (c).

Figure 2. Experimental results of different RGB-T tracking structures on RGBT234 dataset [11]. Teacher denotes the two-stream structure with complex fusion modules. Student denotes a single-stream structure with simple fusion operations. The teacher model employs ResNet50 [7] for feature extraction and fusion modules in [35] for multi-modal feature fusion, respectively. The student model employs ResNet18 [7] for feature extraction and concatenation for multi-modal feature fusion, respectively.

As pointed out by many previous works [15, 19, 20, 29] have made considerable progress on knowledge distillation in multi-modal tasks, they fail to conduct a deep investigation on the huge feature differences between teacher and student in the unimodal feature extraction stage as well as in the multi-modal feature fusion stage, thereby resulting in suboptimal efficiency of the knowledge transformation. For that, a novel teacher-student knowledge distillation training framework, named Cross-Modality Distillation (CMD), is proposed to elaborately guide efficient imitation from three stages: unimodal feature extraction, multi-modal feature fusion and target estimate estimation, as shown in Fig. 1 (c).

Specifically, in the stage of unimodal feature extraction, as pointed out by many previous works [9, 17], the shallower layers of unimodal features usually contain abundant low-level spatial details, which are usually modality-dependent. Differently, the deeper layers of unimodal features often contain many high-level semantic cues, which tend to be strongly modality-consistent. The student model uses a compact single-stream network to extract both RGB features and TIR features, which not only lacks the ability to extract modality-specific information in the shallower layers, but also insufficiently explores the modality-common information in the deeper layers. These interesting observations inspire us to design a Specific-common Feature Distillation (SCFD) module, which transforms the modality-specific information as well as the modality-common information from a two-stream deeper network to a single-stream shallower network.

Second, in the stage of multi-modal feature fusion, the complex multi-modal feature fusion modules in the teacher model show great advantages in various scenarios, while the simple fusion strategies in the student model are usually effective in some specific scenarios. It is difficult for a student model with a single simple fusion strategy to learn more effective complementary information mining capabilities from a complex teacher model due to the large feature differences. Therefore, we design a fusion module with multiple simple fusion strategies in the student model, denoted as Multi-path Selection Distillation (MPSD) module. In the process of learning from the teacher model, the student model can adaptively combine different types of fusion features to make up for the lack of complementary information mining capabilities of a single simple fusion strategy.

Finally, in the stage of target state estimation, with the weakening of the feature representation ability of the student model, the discriminative ability of the tracker for distractors is also reduced. For that, we further present a Hard-focused Response Distillation (HFDR) module to improve the student model’s discriminative ability by alleviating the problem of data imbalance between the targets and the backgrounds, which employs the response maps generated by the teacher model to instruct the student to focus on distinguishing targets from hard negative samples.

As shown in Fig. 2, each of our proposed modules continuously reduces the performance gap between the student model and the teacher model without increasing the number of parameters obviously. To sum up, our work improves an RGB-T tracker dramatically because of the following two contributions:

- A Cross-Modality Distillation (CMD) framework is presented to bridge the performance gap between a compact student model and a powerful teacher model through three stages, i.e., unimodal feature extraction, multi-modal feature fusion and target state estimation. To the best of our knowledge, we are the first to introduce knowledge distillation for multi-modal tracking.

- Experimental results show that our proposed approach helps a student model achieve the state-of-the-art performance on the challenging GTOT [10], RGBT234 [11] and LasHer [14], while reducing the number of parameters and computational complexity.

2. Related Work

RGB-T Tracking Methods. The past few years have witnessed the increase of RGB-T tracking algorithms [4, 10, 13, 17, 31, 39]. Among them, numerous RGB-T trackers [6, 17, 30, 39] have been presented based on the MD-Net [18]. For instance, in [17], Li et al. introduced a multi-adapter architecture to learn modality-common, modality-specific and instance-aware target representations, respectively. In [39], Zhu et al. first presented a network to aggregate the features from all of the layers and all of the modalities. After that, these aggregated features were further

- A Cross-Modality Distillation (CMD) framework is presented to bridge the performance gap between a compact student model and a powerful teacher model through three stages, i.e., unimodal feature extraction, multi-modal feature fusion and target state estimation. To the best of our knowledge, we are the first to introduce knowledge distillation for multi-modal tracking.

- Experimental results show that our proposed approach helps a student model achieve the state-of-the-art performance on the challenging GTOT [10], RGBT234 [11] and LasHer [14], while reducing the number of parameters and computational complexity.

2. Related Work

RGB-T Tracking Methods. The past few years have witnessed the increase of RGB-T tracking algorithms [4, 10, 13, 17, 31, 39]. Among them, numerous RGB-T trackers [6, 17, 30, 39] have been presented based on the MD-Net [18]. For instance, in [17], Li et al. introduced a multi-adapter architecture to learn modality-common, modality-specific and instance-aware target representations, respectively. In [39], Zhu et al. first presented a network to aggregate the features from all of the layers and all of the modalities. After that, these aggregated features were further
Knowledge Distillation Methods. Knowledge Distillation (KD) was first proposed by Hinton et al. [5] to pass dark knowledge from complicated teachers to compact students, enabling students to maintain strong performance as teachers. FitNet [21] proves that the semantic information from intermediate features is also helpful to guide the student model. Besides image classification, KD is widely applied to object detection and object tracking tasks. For instance, [2] and [16] used KD to speed up the detection and segmentation networks, respectively. Furthermore, Wang [24] proposed to design a more compact backbone for faster feature extraction in correlation filter-based trackers. Shen et al. [22] used KD to compress deep Siamese-based trackers for high-performance visual tracking.

In addition, KD is also employed for some multi-modal tasks, such as RGB-D salient object detection [19, 20] and RGB-T pedestrian detection [15, 29]. Specifically, in [20, 29], the single-stream feature extractors and the early fusion strategies were employed in their student models. However, both of them only simply employ the distillation loss functions to improve the performance of student models by using fused features or label knowledges of teacher models, and pay less attention to the huge differences between the teacher model and the student model in the unimodal feature extraction stage as well as in the multi-modal feature fusion stage. Differently, in this paper, we aim to narrow the feature differences between the student model and the teacher model by specifically learning strategies at multiple stages.

3. Distilled RGB-T Tracking

Given a powerful teacher model for RGB-T tracking, the proposed CMD framework aims to prompt a more efficient student model to learn from the teacher model. The knowledge from the teacher model is transferred to the student model to mimic more effective feature representation. This section starts with an overview of the proposed CMD framework. Then, we briefly provide an introduction of the employed teacher and student models. Finally, the three proposed knowledge distillation modules (i.e., SCFD, MPSD and HFRD) are described in details.

3.1. Overview

As illustrated in Fig. 3, the proposed CMD framework includes a teacher model, a student model and three knowledge distillation modules. The teacher model takes a pair of RGB-T images as input, and employs a two-stream feature extractor and several complex multi-modal feature fusion modules for unimodal feature extraction and multi-modal fusion, respectively. Finally, the fused features will be fed into the target state estimation module to obtain the final tracking results. Different from the teacher model, the student model uses a single-stream feature extractor and several efficient multi-modal fusion modules. Although the student model has a higher running speed, the simplification of the model inevitably leads to a decrease in tracking performance.

To make up for the huge performance gap between the student model and the teacher model, the proposed CMD framework attempts to coach the learning process of the student model from three stages: unimodal feature extraction, multi-modal feature fusion and target state estimation. Accordingly, in the first stage, by using a proposed SCFD module, the powerful two-stream feature extraction network of the teacher model will transfer such modality-specific information as well as modality-common information into the single-stream network of the student model to enhance its representation ability for unimodal features. In the second stage, we will present an MPSD module to shrink the differences between the fused features obtained by the teacher model and those obtained by the student model via a multi-path optimization strategy. In the third stage, with a proposed HFRD module, we will adopt the response map generated by the teacher model in a form of spatial attention to instruct the student model to focus on the discrimination of difficult samples, thereby improving its discrimination ability. The improvements in the above three stages will effectively narrow the performance gap between the student model and the teacher model, enabling the student model to achieve competitive tracking results with the teacher model but with fewer parameters and higher computational efficiency.
3.2. Teacher and Student Model

In this section, we will describe the architectures of the employed teacher and student models, which are both based on the recent deep RGB tracker DiMP [1]. As shown in Fig. 4, both the teacher model and the student model can be divided into three stages: unimodal feature extraction, multi-modal feature fusion and target state estimation.

**Feature extraction.** In the teacher model, two feature extractors, denoted as $E_{rgb}$ and $E_{tir}$, regard RGB and TIR modalities in parallel. The two feature extractors both adopt ResNet50 [7] as the backbone to extract multi-level RGB and TIR features, as shown in Fig. 4 (a). Differently, in the student model, only one feature extractor, denoted as $E_{stu}$, regards both RGB and TIR modalities simultaneously. As shown in Fig. 4 (b), $E_{stu}$ just adopts ResNet18 [7] as the backbone for simplification. Similar to the original DiMP tracker, in both the teacher model and the student model, we use the features from block3 and block 4 for classification, and those features only from block4 for regression. The extracted RGB and TIR features from the teacher model are denoted as $f_{rgb}^i$ and $f_{tir}^i$, respectively, and the extracted RGB and TIR features from the student model are denoted as $f_{rgb}^{stu}$ and $f_{tir}^{stu}$, respectively, where $i \in \{1, 2, 3, 4\}$ indexes the feature level.

**Multi-modal feature fusion.** By performing the multimodal fusion modules on the 3rd and 4th levels of RGB and TIR features, we obtain the fused features $f_{rgb}^{fus}$ and $f_{tir}^{fus}$ in the teacher model and the fused features $f_{rgb}^{fus}$ and $f_{tir}^{fus}$ in the student model, respectively. Our teacher model employs a Modality Difference Compensation (MDC) module and a Feature Re-selection module (FRS) for multi-modal feature fusion [35]. Differently, our student model utilizes the proposed MPSD modules for multi-modal feature fusion. Details of MPSD will be introduced in Section 3.4.

**Classification and regression.** Finally, these fused features will be fed to the classification and regression heads, which have the same architectures with those in the original DiMP. Especially, in this stage, the student and teacher models both apply the original classification and regression heads in DiMP. We refer readers to [1, 5] for more details.

3.3. Specific-Common Feature Distillation

This section elaborates on the proposed SCFD module for the two-stage unimodal feature distillation, which lets the single-stream feature extraction module in the student model enable to learn the modality-common information as well as the modality-specific information from the teacher model, as shown in Fig. 5 (a).

We first perform cross-modal interaction on the unimodal RGB features and TIR features from the teacher model to highlight the modality-common information and modality-specific information at different layers, respectively, for better guiding the learning of the student model. Specifically, as shown in Fig. 5 (b), given the unimodal features of shallow layers (i.e., $\{f_{rgb}^i|i = 1, 2, 3\}$) and $\{f_{tir}^i|i = 1, 2, 3\}$) from the teacher model, the proposed Specific Enhanced Modules (SEMs) are employed to obtain such modality-interacted features $f_{rgb, e}$ and $f_{tir, e}^i$ ($i = 1, 2, 3$) with more modality-specific information via subtraction and multiplication. Mathematically,

\[
f_{rgb, e}^i = (f_{rgb}^i \ominus f_{tir}^i) \odot (f_{rgb}^i \ominus f_{tir}^i),
\]

\[
f_{tir, e}^i = (f_{tir}^i \odot f_{tir}^i) \oplus (f_{tir}^i \odot f_{tir}^i),
\]

where $\ominus$, $\odot$, and $\oplus$ denote element-wise subtraction, element-wise addition and element-wise multiplication, respectively. $f_{rgb}^i \oplus f_{tir}^i$ reflects the jointly valid information within RGB and TIR features. While, $f_{rgb}^i \odot f_{tir}^i$ represents the modality-specific information of the RGB modality with respect to the TIR modality. Similarly, the modality-specific information of the TIR modality with respect to the RGB modality can be obtained by $f_{tir}^i \oplus f_{rgb}^i$. Accordingly, $f_{rgb, e}^i$ and $f_{tir, e}^i$ highlight such modality-specific information in addition to preserve jointly valid information, which can be applied to guide feature learning of the student model in shallow layers.

Alternatively, for the RGB and TIR features of deep lay-
ers (i.e., $f_{rgb}^i$ and $f_{tir}^i$), the proposed Consistence Enhanced Module (CEM) is employed to obtain modality-interacted features $f_c^i$ with more modality-common information via addition and multiplication, as shown in Fig. 5 (c). Mathematically,

$$f_c^i = (f_{rgb}^i \oplus f_{tir}^i) \oplus (f_{rgb}^i \odot f_{tir}^i).$$  

(2)

Here, by applying the element-wise addition on $f_{rgb}^i$ and $f_{tir}^i$, the consistency of high-level semantic cues within multi-modal data can be further enhanced. Therefore, $f_c^i$ can better guide the learning of the student model in deep layer.

With the modality-interacted features from the teacher model, the next step is to adjust the feature-channel dimensions of the student model to be consistent with those of the teacher model. Here, inspired by the idea of Knowledge Review [3], we employ a series of attention based fusion (ABF) modules [3] to adjust the channel dimensions of unimodal features and dynamically aggregate the cross-layer features in the student model. The modified features of the student model from ABFs (i.e., $\{u_{rgb}^i| i = 1, 2, 3, 4\}$) and the modality-interacted features of the teacher model (i.e., $\{e_{rgb}^i, e_{tir}^i, e_{rgb}^c, e_{tir}^c\}$) will be employed to force the student model to mimic the specific and common information from the teacher model via a proposed feature-learning distillation loss $L_{SCFD}$, which is formulated as:

$$L_{SCFD} = L_{spe} + L_{com},$$  

(3)

where $l(*)$ denotes the standard MSE loss used in [21].

3.4. Multi-path Selection Distillation

In order to learn the exploration ability of complementary information from the teacher model more effectively, we design a fusion module by using multiple fusion strategies, denoted as Multi-path Selection Distillation (MPSD) module, in the student model. In the process of learning from the teacher model, the student model can adaptively optimize the paths to reduce feature differences.

Specifically, in the student model, the proposed MPSD module first performs multi-modal feature fusion from three typical perspectives: modality differences, modality commonality and modality complementarity. Given the original RGB features $f_{rgb}^i$ and TIR features $f_{tir}^i$ from the 3rd and 4th layers in the student model, three types of initially fused features $f_{fus,1}^i$, $f_{fus,2}^i$ and $f_{fus,3}^i$ are computed by

$$f_{fus,1}^i = sa(f_{rgb}^i \oplus f_{tir}^i),$$  

$$f_{fus,2}^i = f_{rgb}^i \odot f_{tir}^i,$$  

$$f_{fus,3}^i = f_{rgb}^i \oplus f_{tir}^i.$$

(4)

Here, $sa(*)$ denotes the spatial attention mechanism, which first utilizes a convolution layer of kernel size $1 \times 1$ and a softmax layer to get a two-channel weight map. The two-channel weight map is then split into two reliability weight maps for selecting the RGB features and TIR features, respectively. Mathematically, the self-attention mechanism is expressed by:

$$w_{rgb}^i, w_{tir}^i = \sigma(\text{conv}(\text{cat}(f_{rgb}^i, f_{tir}^i), \theta_1)),$$  

$$f_{fus,1}^i = (f_{rgb}^i \odot w_{rgb}^i) \oplus (f_{tir}^i \odot w_{tir}^i),$$  

(5)

where $\text{cat(*)}$ denotes the concatenation operation and $\text{conv(*, \theta_1)}$ denotes a $1 \times 1$ convolutional layer with its parameters $\theta_1$. $\sigma(*)$ denotes the sigmoid layer. The features $f_{fus,1}^i$ mainly reflect the complementary information within multi-modal data. Features $f_{fus,2}^i$ and $f_{fus,3}^i$ reflect their interacted information and their differential information, respectively.

After that, $f_{fus,1}^i$, $f_{fus,2}^i$ and $f_{fus,3}^i$ are further combined together by a weighted fusion way, i.e.,

$$w_1^i, w_2^i, w_3^i = \text{softmax}((f c(gmp(f_{fus,1}^i, f_{fus,2}^i, f_{fus,3}^i))));$$  

$$f_{fus}^i = (f_{fus,1}^i \odot w_1^i) \oplus (f_{fus,2}^i \odot w_2^i) \oplus (f_{fus,3}^i \odot w_3^i),$$

(6)

where $\text{gmp(*)}$ and $fc(*)$ denote the global max pooling layer and the fully connected layers, respectively. $\text{softmax(*)}$ denote the softmax operation. The featurewise weights $w_1^i, w_2^i, w_3^i$ reflect the importance of different fused features for the current scenario. $\odot$ denotes the broadcasting multiplication operation.

With the fused features $\{f_{fus}^i| i = 3, 4\}$ and $\{f_{fus}^i| i = 3, 4\}$ obtained by the teacher model and the student model, respectively, we calculate the fusion distillation loss $L_f_{fus}$.
between the fused features:
\[ L_{\text{fus}} = l(f^1_{\text{fus}}, f^2_{\text{fus}}) + l(f^3_{\text{fus}}, f^4_{\text{fus}}) \] (7)

What’s more, in order to enable the student model to adaptively select a fusion path that is more similar to the teacher model in different scenarios, we introduce an additional penalty \( L_p \) for the efficiency of knowledge transformation during training. More specifically, we first select the fusion type with the smallest difference between the initially fused features of the student model and the fused features of the teacher model by,
\[ L^i_{\text{fus,n}} = l(f^i_{\text{fus}}, f^i_{\text{fus,n}}), n = 1, 2, 3 \]
\[ L^i_{\text{fus,λv}} = \min(L^i_{\text{fus,1}}, L^i_{\text{fus,2}}, L^i_{\text{fus,3}}), \] (8)

where \( λ^1 = 1, 2 \) or 3 denotes the selected type of initially fused features according to the fused feature difference between the teacher and student models.

After that, through the adaptive selection part in MPSD, the student model itself will also predict a type of initially fused features that is suitable for the current tracking scene, i.e.,
\[ w^i_{pr} = \max(w^1_{pr}, w^2_{pr}, w^3_{pr}) \]
where \( ν^i = 1, 2 \) or 3 denotes the predicted type of initially fused features from the student model.

With \( w^i_{pr} \) and \( w^i_{wv} \), we can use a penalty to help the student model choose a fusion path that is more suitable for the current scene under the guidance of the teacher model, i.e.,
\[ L_p = \sum_{i=3}^4 \max(|L^i_{\text{fus,λv}} w^i_{pr} - L^i_{\text{fus,λv}} w^i_{wv}|, 0). \] (10)

By minimizing \( L_p \), \( w^i_{pr} \) and \( w^i_{wv} \), will tend to be consistent, which can enable the student model to adaptively select the fusion path according to the teacher model to improve the exploration ability of complementary information.

On top of that, the overall distill loss in the multi-modal fusion stage can be obtained by:
\[ L_{\text{MPSD}} = L_{\text{fus}} + L_p. \] (11)

### 3.5. Hard-focused Response Distillation

To alleviate the data imbalance problem, we propose the Hard-focused Response Distillation (HFRD) module to instruct the student to focus on distinguishing targets from hard negative samples.

First, we obtain the response map \( R_c ∈ \mathbb{R}^{H×W} \) from the teacher model. Then, in order to prevent the teacher model from failing to have high responses in the target area within some scenes, we use the Gaussian-shaped mask \( R_g ∈ \mathbb{R}^{H×W} \) constructed by ground-truth bounding box as in [1] to correct the response map of the teacher model \( R_c \) as follows:
\[ R_c(i, j) = \begin{cases} R_c(i, j) + R_g(i, j), & \text{if}(R_c(i, j) + R_g(i, j)) < 1, \\ R_g(i, j), & \text{if}(R_c(i, j) + R_g(i, j)) \geq 1. \end{cases} \] (12)

where \( i, j \) are the the horizontal and vertical coordinates of the response map, respectively. The corrected mask \( R_c ∈ \mathbb{R}^{H×W} \) has higher response values not only on the positive samples but also on the hard negative samples.

In the training process of the student model, with the assistance of the corrected mask \( R_c \) from the teacher model, the student model can focus more on distinguishing target from hard negative samples by a proposed Hard-focused Response Distillation loss \( L_{\text{HFRD}} \) to alleviate the data imbalance problem:
\[ L_{\text{HFRD}} = r(R_s ∘ R_c, R_g), \] (13)

where \( r(·) \) denotes the \( L_2 \) loss function [5].

### 3.6. Overall loss

The overall distillation loss \( L_{\text{distill}} \) is the sum of \( L_{\text{SCFD}}, L_{\text{MPSD}} \) and \( L_{\text{HFRD}} \). We train the student model with the total loss as follows:
\[ L_{\text{distill}} = α(L_{\text{SCFD}} + L_{\text{MPSD}}) + βL_{\text{HFRD}} + L_{\text{original}}. \] (14)

where \( α \) and \( β \) are hyper-parameters to balance the distillation loss. \( L_{\text{original}} \) is the original loss for tracking as in [35]. The distillation loss \( L_{\text{SCFD}} \) and \( L_{\text{MPSD}} \) are just calculated on feature maps, which can be easily applied to different takers or other multi-modal vision tasks.

### 4. Experiments

Our tracking approach is implemented in Python based on PyTorch. For inference, we test our tracker on a single Nvidia RTX 1080Ti GPU.

### 4.1. Implementation details

**Training Details.** We adopt the training dataset in LasHeR [14], which contains 979 pairs of RGB-T videos, to train the teacher model and student model, respectively. The proposed CMD framework includes two training stages. In the first stage, we train the teacher model as in MFNet [35] and fix its weights after training. In the second stage, the optimization of the student model is jointly supervised by the original tracking loss \( L_{\text{original}} \) as well as the knowledge transfer loss \( L_{\text{SCFD}}, L_{\text{MPSD}} \) and \( L_{\text{HFRD}} \). \( α \) and \( β \) in Eq. 14 are experimentally set to 0.001 and 100, respectively.

**Online Tracking.** In the tracking phase, our method is similar to DiMP [1]. We split tracking into classification and regression subtasks. For classification subtask, we employ data augmentation [1] on the first frame to construct an initial set, which contains 15 initial training samples for initial classification model training. Then the initial classification model is optimized using the augmented training set during tracking. For regression subtask, the same settings as those in [1] are employed.
4.2. Evaluation datasets and metrics

We evaluate our method on three large-scale benchmark datasets, i.e., GTOT [10], RGBT234 [11] and LasHeR [14]. GTOT is the first standard dataset for RGB-T tracking. It contains 50 RGB-T video sequences annotated with seven challenging attributes. RGBT234 [11] contains 234 pairs of RGB-T videos and 12 annotated attributes. LasHeR is currently the largest RGB-T tracking dataset, which consists of 1244 RGB-T videos with more than 730K frame pairs in total. Among them, 245 videos are used as the testing set, and 979 videos are used as the training set. As in [17], we utilize two widely used metrics, i.e., precision rate (PR) and success rate (SR), to evaluate the tracking performance on GTOT and RGBT234. As in [14], we adopt precision rate (PR), normalized precision rate (NPR) and success rate (SR) to evaluate different trackers on LasHeR.

4.3. Ablation Experiments and Analyses

We conduct some ablation studies on RGBT234 [11] to discuss the impacts of different components in our CMD framework.

**Ablation experiments for each module.** To investigate the impact of each component in our proposed CMD, several versions of our proposed method are provided for comparisons. Specifically, ‘Student’ denotes the model that without any knowledge transformation. Here, the proposed SCFD, MPSD and HFRD are employed in the unimodal feature extraction stage, multi-modal feature fusion stage and target state estimation stage, respectively. The quantitative results of these models are shown in Table 1. It can be seen that SCFD, MPSD and HFRD can all improve the performance of the student model. This verifies that each proposed component in CMD can effectively inherit the knowledge learnt from a powerful teacher model to a student model without obvious loss.

**Effectiveness of the proposed SCFD module.** To further verify the effectiveness of the proposed SCFD module, several variants are also compared with our proposed SCFD module. Here, in the unimodal feature extraction stage, ‘AFD’, ‘SED’ and ‘CED’ denote initially fusing the unimodal features of each layer in the teacher model by using the simple element-wise addition operation, the designed SEM module and the designed CEM module, respectively, and such initially fused features are then employed to guide the student model for single-stream structure learning. As well, ABF modules [3] are employed in all of these variants. As shown in Table 2, the proposed SCFD can better exploit the modality-common and modality-specific information from the teacher model.

**Effectiveness of the proposed MPSD module.** As shown in Table 3, several versions of our proposed MPSD module are also conducted to verify its effectiveness. ‘SAF’ denotes a spatial-wise attention based fusion module. ‘CAF’ denotes a channel-wise attention based module. ‘TF’ denotes adopting the same fusion strategy as that in the teacher model [35]. In particular, each layer adopts the same fusion strategy in the student model. It can be seen that the exploitation of the multi-path fusion strategy can well improve the performance of the student model. In addition, the performance gap between ‘Student-MPSD’ and ‘Student-TF’ is much smaller, which indicates that our proposed MPSD module can better mimic the fused features in the teacher model to compensate for the performance penalty from simple fusion operations.

**Teacher-Student knowledge distillation experiments.** Table 4 shows the performance of using some other knowledge distillation methods in the feature extraction and feature fusion stages for comparisons, including KD [8], FitNets [21], ReviewKD [3] and MD [29]. It is observed from Table 4 that the proposed distillation strategy performs the best. Due to the absence of cross-modal interactions, these existing knowledge distillation methods usually achieve some modest performance gains. In addition, we notice that the student model with a single-stream feature extractor performs obviously well than the student model with a two-stream feature extractor after knowledge distillation. This may be due to the fact that the single-stream network can narrow the modality difference to a certain extent and better acquire the knowledge from the teacher model.

4.4. Comparison with the state-of-the-art

To evaluate the superiority of our proposed method, we compare our method with some existing state-of-the-art RGB-T trackers, including MANet [17], DAFNet [6], DAPNet [39], TODA [28], MACNet [30], CAT [12], CEDiMP
Table 4. Ablation study of different knowledge distillation experiments.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Backbone</td>
<td>2 Res18</td>
<td>2 Res18</td>
<td>2 Res18</td>
<td>1 Res18</td>
<td>1 Res18</td>
<td>1 Res18</td>
</tr>
<tr>
<td>PR/SR</td>
<td>74.5/62</td>
<td>73.9/56.0</td>
<td>79.9/56.5</td>
<td>80.5/56.3</td>
<td>80/55.4</td>
<td>82.4/56.4</td>
</tr>
<tr>
<td>FPS</td>
<td>27</td>
<td>27</td>
<td>27</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Params</td>
<td>31.6</td>
<td>31.6</td>
<td>31.6</td>
<td>19.9</td>
<td>19.9</td>
<td>19.9</td>
</tr>
</tbody>
</table>

Table 5. Quantitative comparisons of our method with some state-of-the-arts methods on benchmark datasets. Higher values indicate better performance.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DAFNet+ [10]</td>
<td>2019</td>
<td>76.6</td>
<td>53.7</td>
<td>82.2</td>
<td>70.7</td>
<td>43.1</td>
</tr>
<tr>
<td>MANet+ [17]</td>
<td>2019</td>
<td>77.7</td>
<td>53.9</td>
<td>89.4</td>
<td>72.4</td>
<td>45.7</td>
</tr>
<tr>
<td>DAFNet+ [10]</td>
<td>2019</td>
<td>79.6</td>
<td>54.4</td>
<td>89.1</td>
<td>71.6</td>
<td>44.9</td>
</tr>
<tr>
<td>mDiMP+ [11]</td>
<td>2019</td>
<td>78.6</td>
<td>55.5</td>
<td>83.6</td>
<td>69.7</td>
<td>44.7</td>
</tr>
<tr>
<td>TOS [25]</td>
<td>2019</td>
<td>78.7</td>
<td>54.5</td>
<td>84.3</td>
<td>67.7</td>
<td>-</td>
</tr>
<tr>
<td>MACNet [30]</td>
<td>2019</td>
<td>79.0</td>
<td>55.4</td>
<td>80.0</td>
<td>71.4</td>
<td>48.3</td>
</tr>
<tr>
<td>CAT1 [17]</td>
<td>2020</td>
<td>80.4</td>
<td>56.1</td>
<td>89.8</td>
<td>71.7</td>
<td>45.1</td>
</tr>
<tr>
<td>FaNet [40]</td>
<td>2020</td>
<td>78.7</td>
<td>55.3</td>
<td>81.8</td>
<td>72.6</td>
<td>44.2</td>
</tr>
<tr>
<td>SiamCA [56]</td>
<td>2020</td>
<td>76.0</td>
<td>56.9</td>
<td>87.7</td>
<td>73.2</td>
<td>-</td>
</tr>
<tr>
<td>JMMC [20]</td>
<td>2020</td>
<td>79.0</td>
<td>57.3</td>
<td>90.2</td>
<td>73.2</td>
<td>-</td>
</tr>
<tr>
<td>MANet [17]</td>
<td>2020</td>
<td>80.0</td>
<td>55.4</td>
<td>82.2</td>
<td>70.7</td>
<td>46.7</td>
</tr>
<tr>
<td>CBPNet [27]</td>
<td>2020</td>
<td>79.4</td>
<td>54.1</td>
<td>88.5</td>
<td>71.6</td>
<td>-</td>
</tr>
<tr>
<td>TPNet [41]</td>
<td>2020</td>
<td>80.6</td>
<td>56.0</td>
<td>88.6</td>
<td>73.9</td>
<td>-</td>
</tr>
<tr>
<td>MFNet [30]</td>
<td>2020</td>
<td>79.3</td>
<td>53.5</td>
<td>89.8</td>
<td>70.7</td>
<td>-</td>
</tr>
<tr>
<td>M5LNet [30]</td>
<td>2020</td>
<td>79.5</td>
<td>54.2</td>
<td>89.6</td>
<td>71.0</td>
<td>-</td>
</tr>
<tr>
<td>HMTF [34]</td>
<td>2020</td>
<td>78.8</td>
<td>56.1</td>
<td>91.2</td>
<td>74.9</td>
<td>-</td>
</tr>
<tr>
<td>APFNet [26]</td>
<td>2020</td>
<td>82.7</td>
<td>57.9</td>
<td>90.5</td>
<td>73.9</td>
<td>50.0</td>
</tr>
<tr>
<td>MANS [11]</td>
<td>2019</td>
<td>78.5</td>
<td>55.7</td>
<td>90.0</td>
<td>72.8</td>
<td>-</td>
</tr>
<tr>
<td>DAFNet+ [10]</td>
<td>2019</td>
<td>80.0</td>
<td>54.9</td>
<td>89.6</td>
<td>70.0</td>
<td>48.0</td>
</tr>
<tr>
<td>mDiMP+ [11]</td>
<td>2019</td>
<td>82.4</td>
<td>58.3</td>
<td>87.7</td>
<td>73.1</td>
<td>58.3</td>
</tr>
<tr>
<td>EAttention [40]</td>
<td>2020</td>
<td>79.4</td>
<td>53.9</td>
<td>90.1</td>
<td>72.1</td>
<td>48.2</td>
</tr>
</tbody>
</table>

On RGBT234. From Table 5, we observe that, in addition to the teacher model, our method achieves the best results with 82.4%/58.4% in PR/SR. In particular, our tracker achieves 3.0%/4.5%, 3.8%/2.9%, and 2.4%/3.6% improvements against FANet+ [40], MANet+ [17], and DAFNet+ [6] in PR/SR, respectively. Compared with the original student model, our algorithm achieves 1.1% improvements in success and 0.7% improvements in precision.

On LasHeR. LasHeR [14] is captured from a number of scenes and categories and is highly diverse. A tracker retrained on this dataset usually achieves some improvements. From Table 5, we can also see that, in addition to the teacher model, our tracker still performs the best in terms of all the three metrics with significant performance superiorities on LasHeR. In particular, our tracker achieves 11.0%/11.9% and 10.8%/12.1% improvements against DAFNet+ [6] and FANet+ [40], which are based on MDNet [18]. Compared with mDiMP+ [31], which is based on DiMP [1] and employs two ResNet50 [7] for feature extraction, our PR/SR is 0.7%/0.8% higher than it. This demonstrates that our proposed method can effectively reduce the performance loss caused by parameter reduction.

5. Conclusion

In this paper, a novel teacher-student knowledge distillation training framework is proposed to reduce the performance gap between a powerful teacher model and a compact student model. Specifically, this framework distills the knowledge from a deep two-stream network with complex multi-modal feature fusion modules to a single-stream network with efficient feature fusion modules. By virtue of the proposed SCFD module, the modality-common information as well as the modality-specific information can be transformed from a two-stream network to a single-stream network in the unimodal feature extraction stage, thus enhancing the representations of unimodal features. Besides, by employing the proposed MPSP module, the student model can adaptively combine multiple fused features generated by various simple fusion strategies to explore complementary information from multi-modal data more thoroughly. In addition, an HFRD module is proposed to improve the student model’s discriminative ability against the distractors by alleviating the problem of data imbalance in the target state estimation stage. Experimental results show that our approach helps a student model achieves the state-of-the-art performance while reducing the number of parameters and computational complexity dramatically.

Limitation: The current method dedicated to reducing computational complexity at the stages of unimodal feature extraction and multi-modal feature fusion, but it paid zero effort to improve the efficiency of target stage estimation, which is our future work.

Acknowledgement: This work is supported by the National Natural Science Foundation of China under Grant No. 61773301. It is also supported by the Shaanxi Innovation Team Project under Grant No. 2018TD-012.
References

[9] Nianchang Huang, Qiang Jiao, Qiang Zhang, and Jungong Han. Middle-level feature fusion for lightweight rgb-d salient object detection. IEEE Transactions on Image Processing, 2022. 2


